DOCUMENT RESUME

ED 308 728 HE 021 792

AUTHOR	Blackwell, J. Lloyd, III
TITLE	Teacher Evaluation. A Revised Case Study of a College in The University of North Dakota. North Dakota Economic Studies, Number 50, Revised.
INSTITUTION	North Dakota Univ., Grand Forks. Bureau of Business and Economic Research.
PUB DATE	Jan 89
NOTE	43p.
AVAILABLE FROM	Bureau of Business and Economic Research, University of North Dakota, Grand Forks, ND 58202 (free).
PUB TYPE	Reports - Descriptive (141) -- Statistical Data (110) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*College Faculty; Educational Quality; Faculty Evaluation; Higher Education; Models; Student Attitudes; *Student Evaluation of Teacher Performance; Teacher Effectiveness; *Teacher Evaluation; Teacher Student Relationship
IDENTIFIERS	University of North Dakota

ABSTRACT

A model explaining various influences on general student opinions of faculty teaching performance was developed for a college in the University of North Dakota (UND). The goals were to (1) find a way to remove influences of improper criteria from general measures of teaching performance and (2) determine which differences in adjusted general ratings of faculty were jointly statistically significant. General ratings of teachers were measured by the percentages of students rating them in the highest category on a survey form. A multiple regression model relating general ratings to the responses from questions designed to explain the general ratings was used to determine factors that seemed to influence general ratings. This permitted neutralization of improper influences. Results indicated that at least some ranking is possible within reasonably small groups, but the ability to achieve any rank-ordering diminishes dramatically as group size increases. It proved possible to unambiguously rank teaching performance into three categories within a group of 10 faculty, but it was impossible to do so for any categorization within a group with as many as 25 members. Four appendices include: a sample UND student evaluation form; advanced treatment of theory and methodology; statistical results of Generalized Least Squares (GLS) estimation; and statistical results of White's heteroskedasticity-consistent estimation. Contains 15 references. (Author/SM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

North Dakota Economic Studies, Number 50, Revised

ED308728

TEACHER EVALUATION

A Revised Case Study of a College in The University of North Dakota

By

J. Lloyd Blackwell III

HE021 792

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Lloyd Blackwell III

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Bureau of Business and Economic Research
University of North Dakota
January 1989

BEST COPY AVAILABLE

Bureau of Business and Economic Research
University of North Dakota
North Dakota Economic Studies, Number 50, Revised

TEACHER EVALUATION

A Revised Case Study of a College in The University of North Dakota

by

J. Lloyd Blackwell, III
Economics Department
University of North Dakota

January 1989

ABSTRACT

A model explaining various influences on general student opinions of faculty teaching performance was developed for a college in the University of North Dakota. Effects of some factors considered to improperly influence student's general opinions were then neutralized, so that general opinions reflected only influences considered proper. It was then determined by testing of differences in adjusted general ratings for joint statistical significance that a group of ten faculty could be rank-ordered into three smaller groups, but that no ordering was possible within a group of twenty-five faculty.

TABLE OF CONTENTS

LIST OF TABLES AND CHARTS.....iii

ACKNOWLEDGEMENTS.....iv

I INTRODUCTION.....1

II THEORY AND METHODOLOGY.....4

III STATISTICAL RESULTS.....6

A. Effects of Adjustments to Ratings.....6

B. Multiple Contrast Results.....10

IV SUMMARY AND CONCLUSIONS.....15

APPENDICES.....17

I Sample of UND Student Evaluation Form.....17

II Advanced Treatment of Theory and Methodology.....18

III Statistical Results of GLS Estimation.....29

IV Statistical Results of White's Heteroskedasticity Consistent Estimation.....31

ENDNOTES.....33

BIBLIOGRAPHY.....35

LIST OF TABLES AND CHARTS

Table — Page

I Sample Sizes, Unadjusted Ratings, and Adjusted Ratings....7

II Unadjusted and Adjusted Ranks....9

Chart

I Wald-type Multiple Contrasts of Adjusted Sample Proportions, All Faculty in a Selected College...11

II Wald-type Multiple Contrasts of Adjusted Sample Proportions, Simulated "Department" of Ten Members....13

ACKNOWLEDGMENTS

This study has a long history of cooperation and help from many people. A key figure in facilitating my studies of the student evaluation process has been Alice T. Clark, Vice President for Academic Affairs, who was sympathetic to the project, and who helped me to gain access to necessary student evaluation data. Richard Balsley, Director of Institutional Research, along with Albert Lindem, User Services Consultant (Computer Center) cooperated in organizing the data into a usable form.

I especially want to thank Dr. William T. Roberds, a former student and now Economist, Federal Reserve Bank of Atlanta, who has matured into my major econometric resource. He provided extremely useful and important insights.

I also wish to thank the members of the Bureau of Business and Economic Research Advisory Committee, not just for seeing fit to provide funding which helped in continuation of this project. Constructive criticisms of early drafts by committee members helped me to turn the final product into a form which seems more comprehensible, even to me. Despite my debts to other people, the responsibility of the final product is, of course, mine alone.

TEACHER EVALUATION

A Revised Case Study of a College in The University of North Dakota

I. Introduction

This paper is a substantive revision of my June, 1988 paper.[1] The major revision is methodological, and stems from constructive criticisms offered by Dr. William T. Roberds[2], who observed that the hypotheses tested were non-linear, requiring a very different theoretical approach than that used in my June, 1988 paper. The revised methodology resulted in more efficient estimation than the previous incorrect approach, and is described in Appendix II.

A difference from my June, 1988 paper is motivated by the way I envision the statistical results similar to those found in this study would probably be used in practice. This new interpretation is incorporated into section III.

Readers of this revision will not suffer from not having read the June, 1988 version. Mention of the revisions is largely for the benefit of those who may have seen the earlier version.

This paper is motivated by two observations: that general measures of teaching performance seem to be explained by criteria having nothing to do with teaching, and that comparisons of various faculty are not typically

done in a statistically rigorous manner. Thus, the goals for this paper are removal of "non-teaching" influences from general measures of teaching performance, and statistically rigorous comparisons of various faculty once "non-teaching" influences are removed from the general measures. The possibilities for meeting these goals are explored here using data from a college in the University of North Dakota, but it is expected that what is learned from this study will be applicable to other academic institutions both in North Dakota and elsewhere.

In my December 1981 study, entitled "Reliability and Comparability of Student Evaluation Data," I explored some problems that might be encountered in using student evaluation data to evaluate teaching performance.[3] In that paper, I concluded that the possibilities for statistical comparisons of faculty were not very promising. As this paper will show, however, I now see brighter prospects.

In the December 1981 paper, I concluded that lack of "complete" knowledge of criteria used by students in evaluating faculty might make statistically reliable teacher evaluation very difficult if not impossible.[4] I now believe that systematic use of "unknown" criteria used in evaluating faculty is detectible. I am also of the opinion that unknown influences on general measures of teaching performance should be treated as "non-teaching" influences, which are then easily removed from general measures. With the problem of unknown criteria solved in this way, it then becomes

possible to develop statistically rigorous comparisons of faculty using a multiple contrast procedure.

With some slight modifications, the methodology developed in this paper can also be used to determine which faculty seem to peform best in given courses, and which courses seem best suited to given faculty. It also seems these procedures could have important applications outside of academia.

The next section of this paper contains a brief non-technical discussion of the theory and methodology applied to the problem at hand. The third section contains a presentation and discussion of statistical results, and the last contains a summary and conclusions.

II. Theory and Methodology

This section is a non-technical discussion of the theory and methodology used in this paper, probably essential to an understanding of the sections to follow. It can also be regarded as an introduction to the more technical treatment contained in Appendix II.

The goals of this study required two things:

1. A statistical model explaining how various factors influence general opinions given by students.

2. A methodology for determining whether differences in adjusted general ratings between various faculty were statistically significant.

General ratings of teachers were measured by the percentages of students in their classes rating them in the highest category in response to question 13 on the UND Student Evaluation Form (See Appendix I for a sample form). A multiple regression model relating general ratings to the responses from questions designed to <u>explain</u> the general ratings, plus some additional information, was used to determine the factors that seemed to influence general ratings. The additional information dealt mainly with questions designed to explain the student's general opinion of the <u>course</u>, as opposed to the teacher. Some of the additional information, however, measured the influences of particular professors and departments, taken to represent important, but unknown influences.

Thus, the multiple regression model permitted removal of influences of factors possibly having nothing to do with teaching from general ratings given by students. General ratings adjusted for non-teaching influences are hereinafter referred to as "adjusted ratings."

Statistical significance of the adjusted ratings was determined using multiple contrasts developed from Wald statistics.[5] Statistically significant differences in adjusted student ratings were then used to rank the quality of teaching by various faculty members.

The data for this study were provided by the UND Office of Institutional Research. The data comprised some 8000 observations on undergraduate lecture sections, for the Fall Semester, 1986. It was decided, however, to limit analysis to one college, which shall, of course, be nameless. For the college subjected to analysis, a total of 1951 observations were available.

III. Statistical Results

A. Effects of Adjustments to Ratings

Table I shows the results of adjustments for factors considered to be improper influences, such as course characteristics, on the general measure of teaching effectiveness. Professors in this group are those for whom 10 or more properly completed student evaluation forms were available. Column 1 of Table I is simply an identification number. Column 2 shows the number of observations provided by the professors coded in column 1. Column 3 shows the percentages of students rating these professors in the highest category in question 13 on the student evlauation form. Column 4 shows the percentages adjusted for influences considered as possibly having nothing to do with teaching.

Table I

Sample Sizes, Unadjusted Ratings, and Adjusted Ratings

By Identification Number

Ident.	Samp.	Unadj.	Adj.
1	33	08.63	08.59
2	17	44.79	46.86
3	69	40.93	41.04
4	42	24.08	24.00
5	29	50.55	51.05
6	31	76.26	75.88
7	27	28.84	30.28
8	12	58.13	56.15
9	11	49.73	48.36
10	21	14.61	14.81
11	41	36.58	41.92
12	52	44.51	45.42
13	24	37.50	34.94
14	72	61.11	50.52
15	24	66.93	66.98
16	23	12.00	11.12
17	59	70.68	69.83
18	18	61.11	41.47
19	54	43.62	43.57
20	85	44.80	44.05
21	12	25.49	25.38
22	58	63.64	63.91
23	41	16.60	16.54
24	12	08.33	25.45
25	22	77.27	70.91

SOURCE: Data were provided by the UND Office of Institutional Research. Analysis was done using SAS and GAUSS.

Obviously, for most professors, adjustments for improper influences made little difference; but there were exceptions. The impact of adjustment on professors 11, 14, 18, 24, and 25 was marked. In the case of professor 24, it was startling. Clearly, for some professors, the impact of "improper" variables on general ratings was not neglible.

Table II shows how rankings of the professors changed when adjustments for improper influences were applied. Column 1 of Table II again shows the professor codes. Column 2 shows unadjusted rankings (by the proportions of students giving the highest rating), with a rank of one being the highest, and twenty-five the lowest. Column 3 then shows adjusted rankings. It is noted that adjustment improved the rankings of nine professors, reduced the rankings of 12, and resulted in no change for four. The largest changes were for professors 18 and 24. Professor 18 fell eight positions, and professor 24 moved up six positions.

Table II

Unadjusted and Adjusted Ranks
Ordered by Faculty Identification Number

Ident.	Un.Rank	Adj.Rank
1	24	25
2	12	10
3	15	16
4	20	21
5	9	7
6	2	1
7	18	18
8	8	6
9	10	9
10	22	23
11	17	14
12	13	11
13	16	17
14	6	8
15	4	4
16	23	24
17	3	3
18	7	15
19	14	13
20	11	12
21	19	20
22	5	5
23	21	22
24	25	19
25	1	2

SOURCE: Data were provided by the UND Office of Institutional Research. Analysis was done using SAS and GAUSS.

B. Multiple Contrast Results

Results of the multiple contrasts of adjusted ratings, for all possible differences among the subjects are shown in Charts I and II. Chart I shows the results for a group of 25 faculty, while Chart II shows results for a group of 10 faculty.

In Charts I and II, rows labled "Rank" indicate adjusted ranks, with one again the highest, and twenty-five the lowest. Rows labelled "Idnt" show professor identification numbers corresponding to the adjusted ranks. Rows labelled "Null" indicate whether a statistically significant difference was found, as explained below. For the procedure employed, all contrasts are <u>simultaneously</u> statistically valid with 95% probability.[4]

The key to interpretation of statistical significance in Chart I and Chart II are the lines of asterisks extending under the ranks and identification codes. To interpret results, first locate any particular professor by identification code. Then look straight down to the line of asterisks beginning under the code. If the line extends under the codes of <u>other</u> professors, the professor at the <u>beginning</u> of the line is not statistically discernible from the other professors to his/her right which are also underlined. On the other hand, professors whose codes begin

CHART I

Wald-type Multiple Contrasts of Adjusted Sample Proportions
Twenty-five Selected Faculty in a Selected College

SOURCE: Data were provided by the UND Office of Institutional Research.
Analysis was done using SAS and GAUSS.

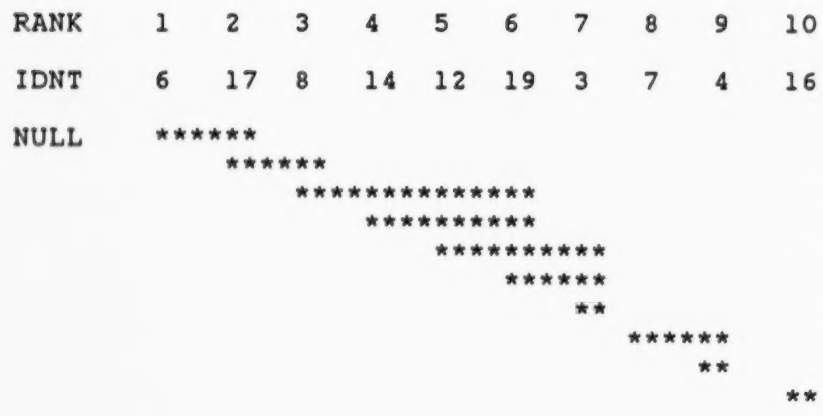
at the left end of a particular line <u>are</u> statistically discernible from professors whose codes are <u>not</u> underlined.

For an example, let us look at Chart I. The highest-ranked professor is number 6; but the line beginning under 6 also extends under 25, 17, 15, 22, 8, 14, 2, 12, and 18. Thus, despite the breaks, professor 6 is indiscernible from 25, 17, 15, 22, 8, 14, 2, 12, and 18, but is discernible from all others. To take another example, professor 3 is statistically indiscernible from 13, 7, 24, 21, and 4, but is statistically discernible from 23, 10, 16, and 1.

Comparing contrasts in Chart II to those in Chart I, it is evident that a greater percentage of multiple contrasts is statistically discernible in the smaller group. Note, for example, that professsor 6 is discernible from 8 and 12 in Chart II, but not Chart I. This is because chi-square statistics underlie both sets of contrasts, and degrees of freedom correspond to group size.[7]

CHART II

Wald-type Multiple Contrasts
Group of Ten Faculty



SOURCE: Data were provided by the UND Office of Institutional Research. Analysis was done using SAS and GAUSS.

At first glance, it appears that the attempt to establish rankings in the group of 25 was reasonably successful, and that the attempt to establish rankings in the group of 10 was much more so. Unfortunately, this may be somewhat of an illusion. To see why, let us look at Chart II, and imagine that we are to allocate salary increases among the group members, using adjusted teacher ratings as the sole criterion.

We begin with professor 6: it is obvious that professor 6 cannot be given a higher raise than professor 17, because

they are statistically indiscernible. But now note that professor 17 is statistically indiscernible from professor 8. Professor 8, in turn, is indiscernible from professors 14, 12, and 19. This chain (a logical syllogism) finally ends with professor 3. It is difficult to argue that professor 6, the highest ranked, should be given a greater salary than professor 3, despite the fact that almost twenty-six percent more students sampled gave professor 6 the highest rating. So professors 6 through professor 3 all get the same raise; but we do have unambiguous support for giving members of this group higher raises than professors 7, 4, and 16. By like reasoning, professors 7 and 4 get higher raises than professor 16. Thus, despite the large percentage of pairwise contrasts which are statistically discernible, it seems unlikely that we can rank faculty into more than three groups, one of which is quite broad. A case can be made that the best we can do is place faculty into three categories.

With less statistical discernability, as in the group of 25 in Chart I, it should be expected that the statistical basis for rank-ordering would be weaker. In fact, if we use the same reasoning applied to Chart II, we must conclude that adjusted student ratings provide no statistically defensible basis for allocating salary raises in the group of 25, which means that all members of the group would have to be given the same raise.

V. Summary and Conclusions.

This paper addressed two goals. The first was to find a way to remove influences of "improper" criteria from general measures of teaching performance. The general measure of teaching performance used in this study was the percentage of students giving the highest possible general rating to evaluated faculty. The second major goal was to determine which differences in adjusted general ratings of faculty were jointly statistically significant.

In order to remove the effects of evaluation criteria deemed improper, a multiple regression model explaining the statistical impact upon the general measure of all available explanatory information was developed. This permitted neutralization of improper influences. In this paper, most criteria assumed to be improper related to information about the course (as opposed to the quality of instruction). Some other "improper" criteria involved student profile information, and statistical evidence of influences shown to be present, but too vague to be considered proper criteria. All information used in obtaining the adjusted general ratings was obtained from UND teacher evaluation forms.

The determination of whether adjusted general measures were statistically significant was based upon multiple contrasts developed from Wald statistics. Results indicated that at least some ranking is possible within reasonably small groups, but that ability to achieve any rank-ordering

whatsoever diminishes dramatically as group size increases. It proved possible to unambiguously rank teaching performance into three categories within a group of 10 faculty, but it was impossible to find unambiguous support for any categorization within a group with as many as 25 members, for which ranking was also attempted, because statistical discernability is related to the number of subjects to be contrasted.'

With one adjustment, the methodology of this paper could be adapted to a second purpose. By making professor/course combinations the units of analysis, one could obtain contrasts of given professors in various courses, and different professors in given courses, as well as different professors regardless of course. This would provide a basis for assigning courses to professors, which the analysis of this paper does not support in its present form.

The basic notion of professor/course analysis may also have application to other than purely academic settings. There is no apparent reason that the methods of this pa-paper, with variables defined to permit analysis of human resource allocation considerations, as discussed immediately above, cannot be applied to any situation where staff members are rated in their performance of several different tasks.

Appendix I. Sample Student Evaluation Form

NCS Trans-Optic® M801-15436-76 A2302

STUDENT EVALUATION FORM

STUDENT ENROLLMENT PROFILE

MAJOR ○ NON MAJOR ○

PERCENT OF CLASSES ATTENDED: 20 ○ 40 ○ 60 ○ 80 ○ 100 ○

GRADE POINT AVERAGE: 4.0 - 3.5 ○ 3.4 - 3.0 ○ 2.9 - 2.5 ○ 2.4 - 2.0 ○ 1.9 - 1.5 ○ 1.4 - 0 ○

WHY ENROLLED: REQ'D COURSE ○ REQ'D AREA ○ ELECTIVE ○

CLASS STATUS: FR ○ SOPH ○ JR ○ SR ○ GRAD ○ SPEC'L ○ MED ○ LAW ○

CALL NBR

CHARACTERISTICS OF THE TEACHER AND TEACHING

	NA	SA	A	N	D	SD
1. The teacher has a thorough grasp of the subject matter.	○	○	○	○	○	○
2. Course objectives and requirements were clearly presented in initial sessions.	○	○	○	○	○	○
3. Course content is organized to meet course objectives.	○	○	○	○	○	○
4. The teacher explains the material clearly.	○	○	○	○	○	○
5. The teacher appears interested in the subject matter of the course.	○	○	○	○	○	○
6. The teacher stimulates enthusiasm for the subject matter of the course.	○	○	○	○	○	○
7. The teacher seems interested in the students' educational progress and needs.	○	○	○	○	○	○
8. The teacher is available for individual help and consultation.	○	○	○	○	○	○
9. The teacher is fair and impartial in dealings with students.	○	○	○	○	○	○
10. The teacher's verbal and non-verbal communications are adequate.	○	○	○	○	○	○
11. The teacher evaluates tests and assignments and communicates results promptly.	○	○	○	○	○	○
12. The teacher gives sufficient notice for required assignments.	○	○	○	○	○	○
13. All things considered, I am favorably impressed by this teacher's instruction.	○	○	○	○	○	○
	○	○	○	○	○	○
	○	○	○	○	○	○

CHARACTERISTICS OF THE LECTURE COURSE
COMPLETE ITEMS 14 - 19 ONLY IF BOX CONTAINS ASTERISK. ☐

	NA	SA	A	N	D	SD
14. The level of the course is appropriate to my ability and experience.	○	○	○	○	○	○
15. The credit given is commensurate with the amount of work required.	○	○	○	○	○	○
16. The required textbooks are helpful and add to the learning experience.	○	○	○	○	○	○
17. Audio-visual and other special aids are utilized effectively.	○	○	○	○	○	○
18. This course is a worthwhile addition to my University experience.	○	○	○	○	○	○
19. In my opinion, the classroom was adequate for this course.	○	○	○	○	○	○
	○	○	○	○	○	○

CHARACTERISTICS OF THE LABORATORY COURSE
COMPLETE ITEMS 20 - 25 ONLY IF BOX CONTAINS ASTERISK. ☐

	NA	SA	A	N	D	SD
20. The laboratory assignments are clear.	○	○	○	○	○	○
21. The laboratory is integrated with the lecture section.	○	○	○	○	○	○
22. The experiments and problems contribute to the learning in the course as a whole.	○	○	○	○	○	○
23. The laboratory space, equipment and supplies are adequate.	○	○	○	○	○	○
24. Opportunity exists for fair and adequate evaluation of laboratory performance.	○	○	○	○	○	○
25. In my opinion, the classroom was adequate for this course.	○	○	○	○	○	○
	○	○	○	○	○	○

CHARACTERISTICS OF THE DISCUSSION COURSE
COMPLETE ITEMS 26 - 31 ONLY IF BOX CONTAINS ASTERISK. ☐

	NA	SA	A	N	D	SD
26. Students are encouraged to question, disagree, and express their own ideas.	○	○	○	○	○	○
27. Assignments and readings are worthwhile.	○	○	○	○	○	○
28. Class discussions contribute to the worth of the course.	○	○	○	○	○	○
29. The credit given is commensurate with the amount of work required.	○	○	○	○	○	○
30. This course is a worthwhile addition to my University experience.	○	○	○	○	○	○
31. In my opinion, the classroom was adequate for this course.	○	○	○	○	○	○
	○	○	○	○	○	○

24

SCALE

NA - NOT APPLICABLE
SA - STRONGLY AGREE
A - AGREE
N - NEUTRAL
D - DISAGREE
SD - STRONGLY DISAGREE

EACH OF THE ITEMS TO THE LEFT DEALS WITH A CHARACTERISTIC OF INSTRUCTORS AND COURSES WHICH STUDENTS FEEL ARE IMPORTANT. INDICATE YOUR RATING OF YOUR INSTRUCTOR AND COURSE BY FILLING IN COMPLETELY ONE OF THE CIRCLES AT THE APPROPRIATE POINT USING THE SCALE ABOVE.

"USE #2 PENCIL"
"ERASE COMPLETELY"
"DO NOT MAKE STRAY MARKS"

TYPE OF COURSE: DISCUSSION ○ LAB ○ LECTURE ○

DO NOT WRITE OUTSIDE THIS BOX.

IF THE INSTRUCTOR'S NAME IS INCORRECT, PLACE THE CORRECT NAME BELOW.

SPACE FOR YOU TO WRITE COMMENTS ON THE OTHER SIDE

APPENDIX II. Advanced Treatment of Theory and Methodology

A. An Overview

The basic methodological tool employed in this study was a multiple regression model used to statistically explain the general measure (GM) of teaching performance. This "GM model" was the reduced-form of a multiple equation model relating the GM to exogenous variables explaining both teacher and course ratings. The GM model was thus directly estimated rather than constructed from estimated structural equations.

Even the GM model, however, was part of a simultaneous system of equations, because it was also neccessary to estimate the means of the exogenous variables for various subjects. Identification of the GM model was achieved by restricting variances and covariances between the GM model and all other equations to zero. With these identifying restrictions, it was possible to estimate the GM model by OLS or GLS.

It was also desirable, as well as reasonable, to apply some restrictions to the models implicitly used in obtaining exogenous variable means for various subjects. Specifically, the means across subjects were assumed to be independent, although the means for particular subjects were not. These restrictions permitted multiple contrasts to be pairwise computed, although all possible pairwise contrasts were simultaneously valid.

Since the student survey forms used Likert scales, all variables in the GM model were discrete, and most variables had five choice categories. Choices for the dependent (endogenous) variable were clearly ordered, however, so it was decided to dichotomize the dependent variable at a "threshold" value. The resulting model was thus of the relatively simple bivariate discrete dependent variable form, in which a value of unity for the dependent variable indicated the threshold value was met or exceeded.

Independent (exogenous) variables were classified in an analysis-of-covariance (ANOCOVA) fashion.' All variables considered to be proper criteria for explaning the dependent variable were treated as covariates, and other variables, the "improper" criteria shown to explain the dependent variable, were treated as concomitants. The model thus permitted contruction of GM's, for each subject, free of the effects of concomitants. This was done by using the covariate means for each subject as predictors for "proper" criteria, but using the means common to all subjects as predictors for concomitants. Thus, the impact of the concomitants upon the general measure was neutralized, permitting construction of adjusted general measures free of concomitant influence.

The adjusted general measures were interpreted, for each subject, as the proportion of their students placing them at or above the threshold value, employing only "proper" criteria.

Once the problem of removal of concomitant impact was solved, the remaining task was to determine which differences in adjusted GM's were simultaneously statistically significant. Because the joint hypotheses were non-linear, owing to the fact that they required estimation of coefficients of the reduced model described above as well as predictors for adusted GM's, multiple contrast procedures designed for linear models were not usable. Hence, hypothesis testing was by means of Wald statistics developed from confidence ellipsoids appropriate to the nonlinear hypotheses. Once the joint significance of the multiple contrasts was established, rankings of subjects found to be pairwise simultaneously statistically discernible from other subjects was possible.

B. The Multiple Regression Model and Methodology

The GM model was specified in the linear form

(1) $y = [X_v | X_c]\beta + u$

where: y = the dichotomous dependent variable.

X_v = the sub-matrix of covariates.

X_c = the sub-matrix of concomitants.

β = the linear regression coefficients.

u = the error vector.

The two problems in models with Bernoulli (binary) dependent variables are, first, that errors are non-normal, and second, that errors are heteroscedastic.[10] The

heteroskedasticity can be handled by using either Generalized Least Squares (GLS), or White's heteroskedasticity-consistent estimation[11]. Non-normality is a problem only in small samples.[12]

In this paper, GLS was used for variable selection and model refinement, and White's heteroskedasticity-consistent estimation was used for final model estimation. The reason for OLS estimation of the final model was that models estimated by OLS have been found to have better prediction properties than GLS models, despite the relative efficiency of GLS.[13] However, GLS was much easier to use in the early stages of model development.

The non-normality of sample regression errors forced reliance upon asymptotic properties of GLS and OLS estimators. The asymptotic normality of $\sqrt{n}(b - \beta)$, for both GLS and OLS estimators, provided the necessary basis for statistical inference.[14] The consistency of the sample mean-squared-error as an estimate of the population error variance[15] then added a basis for joint (Chi-square) tests on all coefficients. Joint tests on selected groups of coefficients (which, it developed, were not neccessary) could have been done with appropriate Wald tests.[16] Since the sample size was 1951, there was no concern that the asymptotic properties neccessary for statistical inference might not be valid.

Except for situations in which there was collinearity among covariates and concomitants, the criterion for

selection of predictors was that Z-statistics for the coefficient should be 1.0 or more.[17] Stein and Ridge techniques were eschewed for variable selection because these usually lead to modifications of coefficients[18], considered unacceptable for this study.

Multicollinearity (collinearity) was particularly severe among the regressors in this problem. Since the intended use for the regression model was essentially prediction, however, this could be a problem only if the process of selecting predictors caused covariates to assume roles of concomitants, or vice-versa. This could happen, for example, if a covariate not meeting the Z = 1.0 criterion for selection was collinear with a concomitant which did. Thus, the impact of the unselected covariate would be captured by the included concomitant, and adjustment for concomitant influences would then to some extent reflect the influence of a covariate.[19]

To avoid the problem of confusing covariates and concomitants, it was clearly necessary to insure that covariates collinear with concomitants, or vice-versa, were retained in the model until all other variables were selected. Once other variables were selected, it would then be possible to conduct joint (Wald) tests on collinear groups containing both covariates and concomitants, if no members of the collinear group met the Z criterion. Of course, if any member of such a collinear group met

individual predictor retention criteria, all members of that group would have to be retained.

Belsley-Kuh-Welsch Collinearity Diagnostics,[10] in the form available in the Statistical Analysis System (SAS)[11] was used to determine collinearity patterns. It was revealed that almost all collinearity was confined to either covariates, or concomitants, but not the two together. One auxilliary regression[12] revealed the possibility of a single instance of pairwise collinearity between a covariate and a concomitant; but both these variables individually met the retention criterion, so both were retained in the model. Hence, there was no concern about confusion of covariates and concomitants in the final model.

B. The Prediction Model.

Adjusted GM for each subject[3] was obtained from the coefficients of the multiple regression model by using

$$(2)\quad m_i = [p_i^v \mid p^c]\begin{bmatrix} b_v \\ b_c \end{bmatrix}$$

where: p_i^v = the vector of sample proportions for the covariates of the i-th subject.

p^c = the vector of sample proportions for the concomitants, common to all subjects.

b_v = the vector of multiple regression coefficients for the covariates.

b_c = the vector of multiple regression coefficients for the concomitants.

m_i = the adjusted GM for the i-th subject (the proportion of students indicating the subject met or exceeded the threshold value, adjusted for concomitant influences.)

C. The Multiple Contrasts.

To easily test the large numbers of differences in adjusted sample proportions for significance, the adjusted proportions were first ranked from highest to lowest. Thus, by subtracting lower-ranked proportions from higher, all differences were positive. Particular contrasts were then statistically significant only when positive differences in adusted sample proportions were significantly greater than contrast variance.

The null hypotheses (contrasts) required to be simultaneously tested were

(2) $H_0 : p_i' b_v - p_j' b_v = 0$,

$i, j = 1, 2, \ldots, q;\ i > j$.

where: p_i' = the vector of predictors for the covariates of the i-th subject.

p_j' = the vector of predictors for the covariates of the j-th subject.

b_v = the vector of coefficients for the covariates of the GM model.

q = the number of subjects to be pairwise contrasted.

Since p^e was the appropriate predictor for concomitants for both subjects in a contrast, $p^e b_e$ was the contribution of the concomitants to the adjusted GM's; hence concomitant influences subtract-out of the nulls.

The variance for (2) can be shown to be

(3) $var(C_{i,j}) = b_v' \cdot var(p_i') \cdot b_v + b_v' \cdot var(p_j') \cdot b_v$

$+ (p_i' - p_j')' \cdot var(b_v) \cdot (p_i' - p_j')$,

where: $C_{i,j}$ = the contrast of subject i to subject j.

$var(p_i')$ = the variance-covariance matrix of the predictors for the i-th subject.

$var(p_j')$ = the variance-covariance matrix of the predictors for the j-th subject.

$var(b_v)$ = the sub-matrix of var(b) corresponding to the covariates.

Since all $C_{i,j}>0$ because of the rankings mentioned above, multiple contrasts were statistically significant iff

$$(4)\quad MC_{i,j} = (p_i' b_v - p_j' b_v) - \sqrt{X^2_{\alpha,q-1} \cdot var(C_{i,j})} > 0 \ .$$

where: $X^2_{\alpha,q-1}$ = a chi-square statistic for significance level α, and q-1 degrees of freedom.

q = the number of subjects among which all possible combinations of multiple contrasts are to be developed.

The total number of pairwise contrasts in a group of size q is the number of pairwise combinations.

D. Variable selection for the GM model

Variable selection was done by GLS, using SAS. This first involved an OLS estimation, in order to obtain weights for weighted least squares, followed by backward elimination to select regressors meeting the selection criterion of Z=1.0 . Two collinear variables, however, were forced into the model, as described above under the collinearity discussion.

Once a set of variables was selected by the above process, another OLS regression was done using the variables selected after the earlier round, the variables re-weighted, and the stepwise process repeated. This selection process was continued until further re-weighting proved pointless. A

White Heteroskedasticity-consistent estimation was then done using the selected variables.

Final results of the variable selection process are shown in Appendix III and Appendix IV. The variables selected represent key classes from all questions on the student evaluation forms, and also the dummy variables representing one department and ten individual professors.

E. The Predictive Model Approach to ANOCOVA.

Readers familiar with traditional ANOCOVA, in which covariates are simply dummy variables which serve to discriminate among subjects, may wonder about the substitution of "teaching" variables for the traditional covariates. The reason is that the "teaching" criteria are supposed to explain the GM. If the traditional approach had been used, the "professor" dummies would have been (as covariates in traditional models always are) "catch-all" dummies for anything and everything that related to the professors, and some of those things may not have been "proper" criteria. Only by expressing the dependent variable in terms of "proper" and "improper" criteria, and then neutralizing the effects of improper criteria, could there be assurance that proper and improper criteria were not confused. Since the proper criteria then substitute for the traditional covariates, if "professor," or "department" dummies are significant in addition to the explanations

provided by proper criteria, they represent unknown influences. Thus, "professor" and "department" dummies were treated as concomitants.

APPENDIX III. Statistical Results of GLS Estimation.

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE
MODEL	49	5340.17455	108.98315	148.832
ERROR	1407	1030.28553	0.73225695	
C TOTAL	1456	6370.46008		

ROOT MSE	0.8557201	R-SQUARE	0.8383
DEP MEAN	0.2902519	ADJ R-SQ	0.8326

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0:
INTERCEP	1	0.08978494	0.11578741	0.775
COV3	1	0.09920281	0.02664271	3.723
COV5	1	0.08214099	0.01651563	4.974
COV8	1	0.03708077	0.01941926	1.909
COV9	1	0.04499300	0.02402623	1.873
COV11	1	-0.06740562	0.02839316	-2.374
COV12	1	-0.05522076	0.01824317	-3.027
COV16	1	0.11485123	0.01741866	6.594
COV17	1	0.32081867	0.02870984	11.175
COV18	1	-0.11715370	0.09967080	-1.175
COV19	1	-0.08605606	0.03075476	-2.798
COV20	1	-0.07344470	0.01710366	-4.294
COV22	1	-0.09289507	0.02993211	-3.104
COV23	1	-0.08645441	0.01740793	-4.966
COV25	1	0.09307248	0.02130006	4.370
COV27	1	-0.05056981	0.01627438	-3.107
COV29	1	0.09638620	0.01880858	5.125
COV30	1	0.04322725	0.03748788	1.153
COV31	1	-0.04357856	0.01537916	-2.834
COV37	1	0.17561667	0.02479671	7.082
COV38	1	-0.11853081	0.03866389	-3.066
COV41	1	0.16177253	0.02023814	7.993
P2	1	0.29859442	0.10548265	2.831
P4	1	0.11307609	0.05266245	2.147
P5	1	0.11792817	0.05224714	2.257
CON3	1	-0.03755317	0.01473428	-2.549
CON5	1	-0.13401789	0.04884938	-2.743
CON6	1	-0.18160440	0.04831717	-3.759
CON7	1	-0.16056465	0.04342220	-3.698

APPENDIX III (Cont'd)

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0:
CON8	1	-0.18336520	0.04609008	-3.978
CON9	1	-0.07297512	0.02517725	-2.898
CON10	1	-0.02979403	0.01809733	-1.646
CON13	1	-0.17650680	0.07917885	-2.229
CON14	1	-0.15042148	0.07465605	-2.015
CON15	1	-0.14982557	0.07396414	-2.026
CON16	1	-0.15911514	0.07524551	-2.115
CON17	1	0.18440922	0.07444224	2.477
CON18	1	0.19992013	0.07513556	2.661
CON19	1	0.20719536	0.07157271	2.895
CON20	1	0.27639321	0.07327409	3.772
D2	1	-0.03515798	0.01713626	-2.052
F12	1	-0.12339508	0.03616910	-3.412
F17	1	0.13656669	0.06965430	1.961
F18	1	0.13560344	0.04819512	2.814
F23	1	0.37940430	0.10369878	3.659
F31	1	-0.24693019	0.08258444	-2.990
F32	1	0.08376783	0.06458480	1.297
F37	1	-0.27269673	0.10533429	-2.589
F50	1	0.11853611	0.04411155	2.687
F51	1	0.11027334	0.07837917	1.407

SOURCE: Data were provided by the UND Office of Institutional Research. Analysis was done using SAS.

APPENDIX IV. Statistical Results of White's Heteroskedasticity Consistent Estimation.

White's Heteroskedasticity Consistent Results

Chi-square = 125.37612 with DF=49
R Sq= 0.76369
Adj R Sq= 0.75760

b est	Std. Error	Z-value
0.10466	0.12179	0.85937
0.05122	0.01888	2.71264
0.10116	0.02022	5.00334
0.03317	0.01955	1.69729
0.01425	0.02732	0.52170
-0.04432	0.02782	-1.59320
-0.04719	0.02218	-2.12775
0.08023	0.01893	4.23841
0.29605	0.03272	9.04790
-0.06523	0.03776	-1.72737
-0.03744	0.02739	-1.36666
-0.04847	0.02210	-2.19310
-0.05564	0.02274	-2.44679
-0.06213	0.01872	-3.31987
0.09803	0.02634	3.72236
-0.04305	0.01842	-2.33727
0.10133	0.02321	4.36625
0.03590	0.02969	1.20937
-0.04428	0.01863	-2.37737
0.16700	0.02809	5.94621
-0.08489	0.03017	-2.81363
0.17076	0.02289	7.46103
0.20448	0.14863	1.37569
0.08654	0.04318	2.00436
0.08733	0.04242	2.05880
-0.02465	0.01849	-1.33326
-0.09626	0.05458	-1.76383
-0.12091	0.05377	-2.24852
-0.11392	0.05087	-2.23939
-0.13451	0.05373	-2.50359
-0.05715	0.02304	-2.48062
-0.02996	0.02013	-1.48857
-0.11669	0.09162	-1.27364
-0.11699	0.09032	-1.29530
-0.11764	0.08976	-1.31058
-0.11876	0.09164	-1.29601
0.12273	0.06840	1.79425
0.14167	0.06722	2.10744
0.13407	0.06272	2.13757

APPENDIX IV (Cont'd)

b est	Std. Error	Z-value
0.17888	0.06613	2.70485
-0.02537	0.02074	-1.22360
-0.05709	0.05689	-1.00354
0.08821	0.08857	0.99590
0.10067	0.03613	2.78677
0.21520	0.09795	2.19704
-0.19355	0.05813	-3.32963
0.05713	0.06231	0.91686
-0.17504	0.13654	-1.28193
-0.05278	0.07409	-0.71232
0.09668	0.08468	1.14171

SOURCE: Data were provided by the UND Office of Institutional Research. Analysis was done using SAS and GAUSS.

NOTE: All variables are listed in same order as in Appendix III.

ENDNOTES

[1] J. Lloyd Blackwell, III, Teacher Evaluation: A Case Study of a College in the University of North Dakota, North Dakota Economic Studies, No. 50 (Grand Forks, North Dakota: Bureau of Business and Economic Research, University of North Dakota, 1988).

[2] Economist, Federal Reserve Bank of Atlanta.

[3] J. Lloyd Blackwell, III, Reliability and Comparability of Student Evaluation Data, North Dakota Economic Studies, No. 32 (Grand Forks, North Dakota: Bureau of Business and Economic Research, University of North Dakota, 1981).

[4] Ibid.

[5] Halbert White, Asymptotic Theory for Econometricians (Orlando, Fla: Academic Press, 1984), chap. 4.

[6] Ibid.

[7] See John Neter, William Wassermann, and Michael H. Kutner, Applied Linear Statistical Models (Homewood, Ill: Richard D. Irwin, Inc., 1985), p. 580. For "q" subjects, only q-1 contrasts will be independent, hence there are only q-1 degrees of freedom in a large-sample context. (The F statistic converges to a chi-square in large samples).

[8] As shown in Appendix II, multiple contrasts involve chi-square statistics, which are directly related to degrees of freedom, which in this case is the number of subjects less one. Larger chi-squares permit less statistical discernability.

[9] This was true in the sense that influences it was desired to control were considered concomitants, as in all ANOCOVA models. Remaining variables were thus covariates, by default. For an introduction to ANOCOVA, see Bradley E. Huitema, The Analysis of Covariance and Alternatives (New York: John Wiley and Sons, 1980), chap. 2.

[10] See Robert S. Pindyck and Daniel L. Rubinfeld, Econometric Models and Economic Forecasts, (2nd ed. ;New York: McGraw-Hill, 1981), pp. 274-280. For more advanced references, see two articles by S. Warner, "Asymptotic Variances for Dummy Variable Regression Under Normality Assumptions," Journal of the American Statistical Association, 62 (1967): 1305-1314, and "Multivariate Regression of Dummy Variates under Normality Assumptions," Journal of the American Statistical Association, 58 (1963): 1054-1063.

[11] Halbert White, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test For Heteroskedasticity," Econometrica, 48, No. 4 (May, 1980), 817-838.

[12] Henri Theil, Principles of Econometrics (New York: John Wiley and Sons, Inc., 1971), chap. 6, pp. 380-381.

[13] T. Domincich and D. McFadden, Urban Travel Demand - A Behavioral Analysis, (Amsterdam: North Holland, 1975), chap. 5.

[14] Theil, Econometrics, pp. 380-381, 403-405.

[15] Ibid., p. 380.

[16] George G. Judge, William E. Griffiths, R. Carter Hill, and Tsoung-Chao Lee, The Theory and Practice of Econometrics (New York: John Wiley and Sons, Inc., 1980), p. 757.

[17] Ibid., pp. 417-422.

[18] Ibid. pp. 428, 472.

[19] J. Johnston, Econometrics (2nd. ed. ;New York: McGraw-Hill, 1963), pp. 168-169.

[20] David A. Belsley, Edwin Kuh, and Roy E. Welsch, Collinearity Diagnostics (New York: John Wiley and Sons, Inc., 1980), pp. 106-107.

[21] SAS Institute,The SAS Users Guide: Statistics (Version 5 Ed.; Cary, N.C.: SAS Institute, Inc., 1985), chap. 31.

[22] Belsley, Kuh, and Welsch, Collinearity Diagnostics, pp. 112-113.

[23] Huitema, The Analysis of Covariance and Alternatives, pp. 31-38, 65-66.

BIBLIOGRAPHY

Belsley, David A.; Kuh, Edwin; Welsch, Roy E. Collinearity Diagnostics. New York: John Wiley and Sons, Inc., 1980.

Blackwell, III, J. Lloyd . Reliability and Comparability of Student Evaluation Data. North Dakota Economic Studies, No. 32. Grand Forks, North Dakota: Bureau of Business and Economic Research, University of North Dakota, 1981.

Blackwell, III, J. Lloyd . Teacher Evaluation: A Case Study of a College in the University of North Dakota. North Dakota Economic Studies, No. 50. Grand Forks, North Dakota: Bureau of Business and Economic Research, University of North Dakota, 1988.

Domincich, T. and McFadden, D. Urban Travel Demand - A Behavioral Analysis. Amsterdam: North Holland, 1975.

Goldberger, Arthur S. Econometric Theory. New York: McGraw-Hill, 1964.

Huitema, Bradley E. The Analysis of Covariance and Alternatives. New York: John Wiley and Sons, 1980.

Johnston, J. Econometrics, 2nd. ed. New York: McGraw-Hill, 1963.

Judge, George G.; Griffiths, William E.; Hill, R. Carter; Lee, Tsoung-Chao. The Theory and Practice of Econometrics. New York: John Wiley and Sons, Inc., 1980.

Neter, John, Wasserman, William, and Kutner, Michael H. Applied Linear Statistical Models, 2nd ed. Homewood, Ill: Richard D. Irwin, Inc., 1985.

Pindyck, Robert S., and Rubinfeld, Daniel L. Econometric Models and Economic Forecasts, 2nd ed. New York: McGraw-Hill, 1981.

SAS Institute. The SAS Users Guide: Statistics, Version 5 Edition. Cary, N.C.: SAS Institute, Inc., 1985.

Theil, Henri. Principles of Econometrics. New York: John Wiley and Sons, Inc., 1971.

Warner, S. "Asymptotic Variances for Dummy Variable Regression Under Normality Assumptions." Journal of the American Statistical Association, 62 (1967): 1305-1314.

----------. "Multivariate Regression of Dummy Variates under Normality Assumptions." Journal of the American Statistical Association, 58 (1963): 1054-1063.

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test For Heteroskedasticity." Econometrica, 48, No. 4 (May, 1980): 817-838.

_____. Asymptotic Theory for Econometricians. Ordando, Fla: Academic Press, 1984), chap. 4.